"Point and Interval Estimation" An Article for the International Encyclopedia of the Social and Behavioral Sciences

George Casella*
Cornell University

Roger L. Berger
North Carolina State University

September 23, 1999

Point Estimation

When sampling is from a population described by a density or mass function $f(x|\theta)$, knowledge of θ yields knowledge of the entire population. Hence, it is natural to seek a method of finding a good estimator of the point θ, that is, a good point estimator.

In many cases, there will be an obvious or natural candidate for a point estimator of a particular parameter. For example, the sample mean is a natural candidate for a point estimator of the population mean. However, when we leave a simple case like this, intuition may desert us so it is useful to have some techniques that will at least give us some reasonable candidates for consideration. Those that have stood the test of time include:

1. *The Method of Moments*

   The method of moments (MOM) is, perhaps, the oldest method of finding point estimators, dating back at least to Karl Pearson in the late 1800s. One of the strengths of MOM estimators is that they are usually simple to use and almost always yields some sort of estimate. In many cases, unfortunately, this method yields estimators that may be improved upon.

   Let $X_1, \ldots, X_n$ be a sample from a population with density or mass function $f(x|\theta_1, \ldots, \theta_k)$. MOM estimators are found by equating the

*Supported by National Science Foundation Grant DMS-9971586. Email: gc15@cornell.edu. This is technical report BU-1455-M in the Department of Biometrics, Cornell University, Ithaca, NY 14853.

first k sample moments to the corresponding k population moments. That is, we define the sample moments by $m_j = \sum_{i=1}^n X_i^j$ and the population moments by $\mu_j(\theta_1, \ldots, \theta_k) = \mathrm{E}X^j$ for $j = 1, \ldots, k$. We then set $m_j = \mu_j(\theta_1, \ldots, \theta_k)$ and solve for $\theta_1, \ldots, \theta_k$. This solution is the MOM estimator of $\theta_1, \ldots, \theta_k$.

2. *Maximum Likelihood Estimators*

For a sample $X_1, \ldots, X_n$ from $f(x|\theta_1, \ldots, \theta_k)$, the likelihood function is defined by

$$L(\theta|\mathbf{x}) = L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) = \prod_{i=1}^n f(x_i|\theta_1, \ldots, \theta_k).$$

The values of θ_i that maximize this function are those parameter values for which the observed sample is most likely, and are called the *maximum likelihood estimators (MLE)*. If the likelihood function is differentiable (in θ_i), the MLEs can often be found by solving

$$\frac{\partial}{\partial \theta_i} \log L(\theta|\mathbf{x}) = 0, \; i = 1, \ldots, k.$$

where the vector with coordinates $\frac{\partial}{\partial \theta_i} \log L(\theta|\mathbf{x})$ is called the *score function* (see Schervish 1995, Section 2.3).

Example If $X_1, \ldots, X_n$ are iid Bernoulli(p), the likelihood function is

$$L(p|\mathbf{x}) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}$$

and differentiating $\log L(p|\mathbf{x})$ and setting the result equal to zero gives the MLE $\hat{p} = \sum_i x_i/n$. This is also the Method of Moments estimator.

If we instead have samples $X_1, \ldots, X_n$ from a binomial(k, p) population where p is known and k is unknown, the likelihood function is

$$L(k|\mathbf{x}, p) = \prod_{i=1}^n \binom{k}{x_i} p^{x_i}(1-p)^{k-x_i}.$$

and the MLE must be found by numerical maximization. The method of moments will give the closed form solution

$$\hat{k} = \frac{\bar{x}^2}{\bar{x} - (1/n)\sum(x_i - \bar{x})^2}$$

which can take on negative values. This illustrates a shortcoming of the method of moments, one not shared by the MLE. Another, perhaps more serious shortcoming of the MOM estimator is that it may not be based on a *sufficient statistic*, which means it could be inefficient in not using all of the available information in a sample. In contrast, both MLEs and Bayes estimators are based on sufficient statistics.

3. *Bayes Estimators*

   In the Bayesian paradigm a random sample $X_1, \ldots, X_n$ is drawn from a population indexed by θ and, where θ is considered to be a quantity whose variation can be described by a probability distribution (called the *prior distribution*). A sample is then taken from a population indexed by θ and the prior distribution is updated with this sample information. The updated prior is called the *posterior distribution.*

   If we denote the prior distribution by $\pi(\theta)$, and the sampling distribution by $f(\mathbf{x}|\theta)$, then the posterior distribution, the conditional distribution of θ given the sample, $\mathbf{x}$, is

$$\pi(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)\pi(\theta)/m(\mathbf{x})$$

   where $m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)d\theta$ is the marginal distribution of $\mathbf{x}$.

   Example Let $X_1, \ldots, X_n$ be iid Bernoulli(p). Then $Y = \sum X_i$ is binomial(n, p). The posterior distribution of p given y, is

$$f(p|y) = \frac{f(y,p)}{f(y)} = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1},$$

   which is a beta distribution with parameters $y+\alpha$ and $n-y+\beta$. The mean, a Bayes estimator of p, is

$$\hat{p}_{\mathrm{B}} = \frac{y+\alpha}{\alpha+\beta+n}.$$

There are many other methods of deriving point estimators (robust methods, least squares, estimating equations, invariance) but the three mentioned above are among the most popular. No matter what method is used to derive a point estimator, it is important to evaluate the estimation using some performance criterion.

We can loosely group evaluation criteria into *large sample* or *asymptotic* methods, and *small sample* methods. In large samples, MLEs typically perform very well, being asymptotically normal and *efficient*, that is, attaining the smallest possible variance. Other types of estimators that are derived in a similar manner (for example *M-estimators*) also share good asymptotic properties. For a detailed discussion see Lehmann (1999) or Lehmann and Casella (1998, Chapter 6).

In small samples, estimators can be evaluated using *mean squared error* (MSE). The MSE of an estimator W of a parameter θ is the function of θ defined by $\mathrm{E}_\theta(W-\theta)^2$. It has the interpretation

$$\mathrm{E}_\theta(W-\theta)^2 = \mathrm{Var}_\theta\, W + (\mathrm{E}_\theta W - \theta)^2 = \mathrm{Var}_\theta\, W + (\mathrm{Bias}_\theta\, W)^2,$$

where the *bias* of an estimator W is $\mathrm{E}_\theta W - \theta$.

Example Under normality, the MLE of the variance σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{n-1}{n}S^2$, where $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ is the usual unbiased estimate. Straightforward calculation gives

$$\mathrm{E}(\hat{\sigma}^2 - \sigma^2)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4 \text{ and } \mathrm{E}(S^2 - \sigma^2)^2 = \left(\frac{2}{n-1}\right)\sigma^4,$$

showing that $\hat{\sigma}^2$ has smaller MSE than S^2. This is an example of a variance/bias trade off, as the biased $\hat{\sigma}^2$ has a smaller variance, resulting in a smaller MSE.

Interval Estimation

Reporting a point estimator of a parameter θ only provides part of the story. The story becomes more complete if an assessment of the error of estimation is also reported. Informally, this can be accomplished by giving an estimated standard error of the estimator and, more formally, this becomes the reporting of an *interval estimate*. If $\mathbf{X} = \mathbf{x}$ is observed, an interval estimate of a parameter θ is a pair of functions, $L(\mathbf{x})$ and $U(\mathbf{x})$ for which the inference $\theta \in [L(\mathbf{x}), U(\mathbf{x})]$ is made. The *coverage probability* of the random interval $[L(\mathbf{X}), U(\mathbf{X})]$ is the probability that $[L(\mathbf{X}), U(\mathbf{X})]$ covers the true parameter, θ, and is denoted by $P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$.

By definition, the coverage probability depends on the unknown θ, so cannot be reported. What is typically reported is the *confidence coefficient*, the infimum of the coverage probabilities, $\inf_\theta P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$.

If $X_1, \ldots, X_n$ are iid with mean μ and variance σ^2, a common interval estimator for μ is

$$\mu \in \bar{x} \pm 2\frac{s}{\sqrt{n}} \tag{1}$$

where $\bar{x}$ is the sample mean and s is the sample standard deviation. The validity of this interval can be justified from the Central Limit Theorem, since

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \to \mathcal{N}(0,1),$$

the standard normal distribution. We then see that the coverage probability (and confidence coefficient) of (1) is approximately 95%.

The above interval is a "large sample" interval since its justification is based on an asymptotic argument. There are many methods for constructing interval estimators that are valid in small samples, of which the following are a sample:

1. *Inverting a Test Statistic*

   There is a correspondence between acceptance regions of tests and confidence sets, summarized in the following theorem.

   Theorem 1 *For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test of $H_0: \theta = \theta_0$. For each $\mathbf{x} \in \mathcal{X}$, define a set $C(\mathbf{x})$ in the parameter space by*

   $$C(\mathbf{x}) = \{\theta_0 ; \mathbf{x} \in A(\theta_0)\}.$$

   Then the random set $C(\mathbf{X})$ is a $1-\alpha$ confidence set. Conversely, let $C(\mathbf{X})$ be a $1-\alpha$ confidence set. For any $\theta_0 \in \Theta$, define

   $$A(\theta_0) = \{\mathbf{x}: \theta_0 \in C(\mathbf{x})\}.$$

   Then $A(\theta_0)$ is the acceptance region of a level α test of $H_0: \theta = \theta_0$.

   Example If $X_1, \ldots, X_n$ are iid $\mathrm{n}(\mu, \sigma^2)$, with σ^2 known, the test of $H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$ will accept the null hypothesis at level α if

   $$\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \square \mu_0 \square \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

   The interval of μ values, $[\bar{x} - z_{\alpha/2}\sigma/\sqrt{n},\ \overline{x} + z_{\alpha/2}\sigma/\sqrt{n}]$, for which the null hypothesis will be accepted at level α, is a $1-\alpha$ confidence interval for μ.

2. *Pivotal Inference*

Perhaps one of the most elegant methods of constructing set estimators is the use of pivotal quantities (Barnard 1949). A random variable $Q(\mathbf{X}, \theta) = Q(X_1, \ldots, X_n, \theta)$, is a *pivotal quantity* (or *pivot*) if the distribution of $Q(\mathbf{X}, \theta)$ is independent of all parameters. If we find a set C such that $P(Q(\mathbf{X}, \theta) \in C) = 1 - \alpha$, then the set $\{\theta : Q(\mathbf{X}, \theta) \in C\}$ has coverage probability $1 - \alpha$.

In location and scale cases, once we calculate the sample mean $\bar{X}$ and the sample standard deviation S, we can construct the following pivots:

Form of pdf	Type of pdf	Pivotal quantity
$f(x-\mu)$	location	$\overline{X} - \mu$
$\frac{1}{\sigma} f(\frac{x}{\sigma})$	scale	$\frac{\overline{X}}{\sigma}$
$\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$	location-scale	$\frac{\overline{X}-\mu}{S}$

In general, *differences* are pivotal for location problems, while *ratios* (or products) are pivotal for scale problems.

Example Suppose that $X_1, \ldots, X_n$ are iid exponential(λ). Then $T = \sum X_i$ is a sufficient statistic for λ and $T \sim$ gamma(n, λ). In the gamma pdf t and λ appear together as t/λ and, in fact the gamma(n, λ) pdf $(\Gamma(n)\lambda^n)^{-1} t^{n-1} e^{-t/\lambda}$ is a scale family. Thus, if $Q(T, \lambda) = 2T/\lambda$, then

$$Q(T, \lambda) \sim \text{gamma}(n, \lambda(2/\lambda)) = \text{gamma}(n, 2),$$

which does not depend on λ. The quantity $Q(T, \lambda) = 2T/\lambda$ is a pivot with a gamma$(n, 2)$, or χ^2_{2n}, distribution, and a $1 - \alpha$ pivotal interval is

$$\frac{2T}{\chi^2_{2n,\alpha/2}} \square \lambda \square \frac{2T}{\chi^2_{2n,1-\alpha/2}},$$

where $P(\chi^2_{2n} > \chi^2_{2n,a}) = a$.

3. *Bayesian Intervals*

If $\pi(\theta|\mathbf{x})$ is the posterior distribution of θ given $\mathbf{X} = \mathbf{x}$, then for any set $A \subset \Theta$ the posterior probability of A is

$$P(\theta \in A|\mathbf{x}) = \int_A \pi(\theta|\mathbf{x})\, d\theta,$$

and A is called a *credible set* for θ. If $\pi(\theta|\mathbf{x})$ is a pmf, we replace integrals with sums in the above expressions.

The interpretation of the Bayes interval estimator is different from the classical intervals. In the classical approach, to assert 95% coverage is to assert that in 95% of repeated experiments, the realized intervals will cover the true parameter. In the Bayesian approach, a 95% coverage means that the probability is 95% that the parameter is in the realized interval. In the classical approach the randomness comes from the repetition of experiments, while in the Bayesian approach the randomness comes from the prior distribution.

Example Let $X_1, \ldots, X_n$ be iid Poisson(λ) and assume that λ has a gamma prior pdf, $\lambda \sim \text{gamma}(a, b)$, where a is an integer. The posterior pdf of λ is

$$\pi(\lambda|\textstyle\sum X = \sum x) = \text{gamma}(a + \sum x, [n + (1/b)]^{-1}).$$

Thus the posterior distribution of $2[n+(1/b)]\lambda$ is $\chi^2_{2(a+\Sigma x)}$, and a $1-\alpha$ Bayes credible interval for λ is

$$\left\{\lambda\colon \frac{\chi^2_{2(a+\Sigma x),1-\alpha/2}}{2[n+(1/b)]} \,\square\, \lambda \,\square\, \frac{\chi^2_{2(a+\Sigma x),\alpha/2}}{2[n+(1/b)]}\right\}.$$

We can also form a Bayes set by taking the *highest posterior density* (HPD) region of the parameter space, by choosing c so that

$$1-\alpha = \int_{\{\lambda:\pi(\lambda|\Sigma x)\geq c\}} \pi(\lambda|\textstyle\sum x)\, d\lambda.$$

Such a construction is optimal in the sense of giving the shortest interval for a given $1-\alpha$ (although if the posterior is multimodal the set may not be an interval).

For more details on constructing and evaluating intervals see Casella and Berger (1990).

References

1. Barnard, G. A. (1949). Statistical Inference (with discussion). *Journal of the Royal Statistical Society, Series B* 11, 115–139.

2. Casella, G. and Berger, R. L. (1990). *Statistical Inference.* Pacific Grove, CA: Wadsworth/Brooks Cole.

3. Lehmann, E. L. (1998). *Introduction to Large-Sample Theory.* New York: Springer-Verlag.

4. Lehmann and Casella (1998). *Theory of Point Estimation, Second Edition.* New York: Springer-Verlag.

5. Schervish, M. (1995). *Theory of Statistics.* New York: Springer-Verlag.